

A7

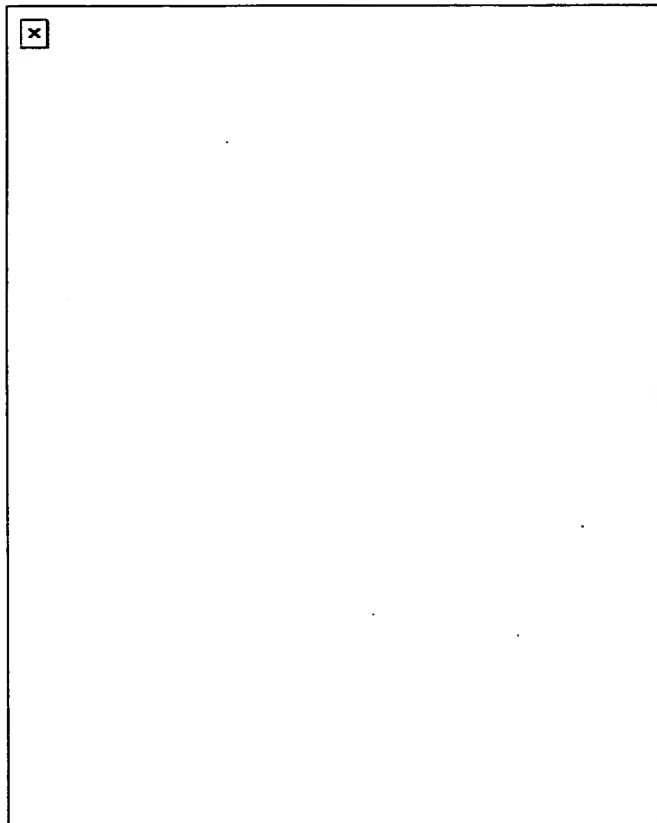
DECISION TREE LEARNING AND GENERATING DEVICE

Patent number: JP7093158
Publication date: 1995-04-07
Inventor: ORIHARA RYOHEI
Applicant: TOSHIBA CORP
Classification:
- International: G06F9/44; G06F15/18
- european:
Application number: JP19930238150 19930924
Priority number(s):

Abstract of JP7093158

PURPOSE:To automatically generate a decision tree for feature amount classification to take continuous values.

CONSTITUTION:As the device to obtain rules for classifying a case data (DT) set into correspondent classification classes (CL) among previously decided CL corresponding to the plural kinds of feature amounts concerning the set of plural DT provided with the two kinds of feature amounts of a first feature amount (D1) expressing the feature of a case and a second feature amount (D2) expressing an attribute for classification at least, this device is constituted by providing a means to successively fetch the plural applied DT, to calculate difference (df) between the maximum value and minimum value of D1 by attributes, to divide it into a certain number of calculate decided by the df and relevant grading until the value of this df after division gets less than the value of grading of classification decided for classifying D1, to define the result as CL and to generate the decision tree by deciding the critical value range of the feature values of respective adjacent CL with the intermediate value of the closest value among the feature values of respective CL in the case of division, and a means to finish the generation of the decision tree when the class division error rate of respective CL reaches a previously decided set value.



Data supplied from the **esp@cenet** database - Worldwide

THIS PAGE BLANK (USPTO)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-93158

(43) 公開日 平成7年(1995)4月7日

(51) IntCl. ⁹	識別記号	序内整理番号	F I	技術表示箇所
G 0 6 F 9/44	5 5 0 M	9193-5B		
15/18	5 6 0 A	9365-5L		

審査請求 未請求 請求項の数 1 O L (全 16 頁)

(21) 出願番号 特願平5-238150

(22) 出願日 平成5年(1993)9月24日

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72) 発明者 折原 良平

神奈川県川崎市幸区柳町70番地 株式会社

東芝柳町工場内

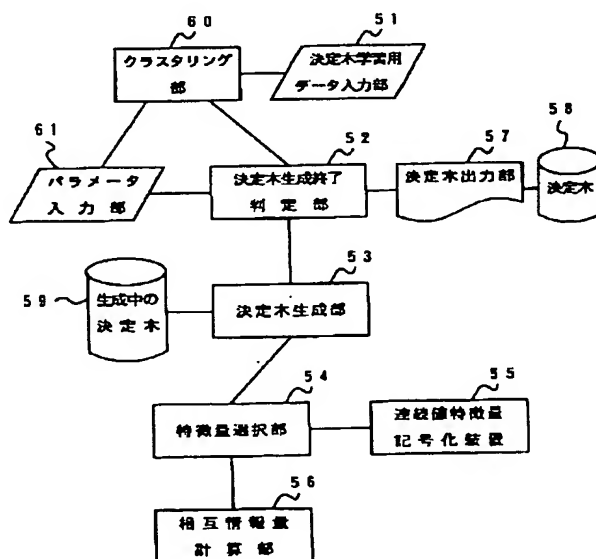
(74) 代理人 弁理士 鈴江 武彦

(54) 【発明の名称】 決定木学習生成装置

(57) 【要約】

【目的】連続値をとる特徴量分類用の決定木を自動生成できるようにする。

【構成】事例の特徴を表す第1特徴量(D1)と、分類用の属性を示す第2特徴量(D2)との少なくとも2種の特徴量を有する事例データ(DT)を対象とし、複数のDTの集合に対し、前記複数種の特徴量に応じてそのDT集合を、予め定めた分類クラス(CL)のうちの対応するCLに分類するための規則を得る装置として、与えられた複数のDTを順次取り込み、そのD1の最大値と最小値との差(df)を属性別に求めこのdfの分割後の値がD1を分類するために定めた分類の粒度の値以下となるまでdfと当該粒度の値とにより定まるクラスタ数分に分割してCLとすると共に、各隣接するCLの特徴値の境界値範囲を前記分割の際の各CLの特徴値のうちの最も近い値の中間値を以て定めることで決定木を生成する手段、CL夫々のCLのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する手段とを設けて構成した。



1

【特許請求の範囲】

【請求項 1】 事例の特徴を表す第 1 の特徴量と、分類用の属性を示す第 2 の特徴量との少なくとも 2 種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習生成装置として、

与えられた複数の事例データを順次取り込み、その第 1 の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第 1 の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラス数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成したことを特徴とする決定木学習生成装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、与えられたいくつかの事例データを、適当な部分クラスに分類する分類規則を、事例データから帰納的に学習する決定木学習装置に関する。

【0002】

【従来の技術】近年、人口知能の研究において、例題を与えることによって、対象とする概念の一般知識を発見する機械学習の技術が広く研究されつつある。その一つの方向として、数値または記号で表現された幾つかの特徴量と、それに対してオペレータ等が事前にデータを吟味して与える分類結果を事例とし、多数の事例を分類装置の例題として与えることによって、特徴量と分類結果の間の一般法則を見付け出し、新たな事例が与えられた時に、その事例の持つ特徴量を評価して分類結果を出力すると云った処理をする学習装置が開発されている。

【0003】ここで、事例が分類されるべき部分クラスの分類結果を分類クラスと呼ぶ。例えば、図 7 に示す事例データは、2 つの特徴量 N_1 ($=604 \sim 695$) および N_2 ($=A, B$) と、分類クラス ($C_1 \sim C_5$) とを持つ事例データである。

【0004】そして、事例データの持つ特徴量を評価してその事例データの分類結果を得ると云ったような分類規則を導く方法の代表例として、特徴量を分類属性とし、分類結果クラスを得る決定木（「特徴量 A の値が a であり、特徴量 B の値が b であれば分類結果は C である」と云った規則を、特徴量を“節”とし、分類結果を“葉”とする木（ツリー）の形にまとめたもの）を導くようにする手法がある。

2

【0005】図 8 は、上述した図 7 の事例に対して、以下で説明する公知の決定木学習アルゴリズムが導いた決定木の例を示した図である。この決定木は、「特徴量 N_2 の値が“A”であって、特徴量 N_1 の値が“614.5”以下ならば、分類クラスは C_4 、そうでなく、特徴量 N_1 の値が“614.5”より大きく“649.5”以下ならば、分類クラスは C_5 、そうでなく…」と云う知識を表す。

【0006】図 9 は、公知の決定木学習装置の構成を示したものであり、図 10 は公知の決定木学習装置の処理の流れを示したものである。図 9 において、91 は決定木学習用データ入力部であって、当該決定木学習装置に決定木学習用データを入力するものであり、92 は決定木生成終了判定部であって、特徴量選択部 94 によって選択された特徴量がどの値であるかによって、与えられたデータ S を複数の部分に分割し、それぞれの部分について図 10 の S t 81 のステップ以降の処理を再帰的に繰り返すことで、全ての枝が終了したか否かを判断するものであり、この時、生成中の決定木 99 に、選択された特徴量を付け加えることで決定木を成長させるのが決定木生成部 93 である。94 は特徴量選択部であって、特徴量がどの値であるかを選択するものである。また、95 は連続値を区間に分割することによって記号の特徴量に変換する連続値特徴量記号化装置、96 は相互情報量計算部であって、与えられたデータ S の持つ情報量 $I(S)$ と、すべての特徴量 A_i について、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和 $E(A_i, S)$ 、さらにその差 $gain(A_i, S) = I(S) - E(A_i, S)$ を計算するものであり、この差 $gain(A_i, S)$ が最も大きい特徴量を選ぶのが上述の特徴量選択部 94 である。

【0007】97 は決定木出力部であって、決定木生成終了判定部 92 が決定木生成終了と判定した場合に、生成中の決定木 99 を完成した決定木として出力するものであり、98 がこの決定木出力部 97 の出力した決定木である。

【0008】このような構成の従来装置は、つぎのような処理を行う。決定木学習用データ入力部 91 から決定木学習用データが入力されると（図 10 の S t 80）、決定木生成終了判定部 92 によって、分類クラスに分けられた事例データが、その分類クラスに該当するかの誤り率である分類誤り率設定値（prune rate）以上を占めるか否かが判断される（図 10 の S t 81）。

【0009】もしそうであって、すべての再帰的繰り返しについて、やはり同じことが言えているならば（図 10 の S t 82）、決定木出力部 97 が生成中の決定木 99 を決定木 98 として出力し（図 10 の S t 83）、処理を終了する。

【0010】そうでないならば、決定木生成部 93 は、特徴量選択部 94 によって選択された特徴量がどの値で

3

あるかによって、与えられたデータSを複数の部分に分割し、それぞれの部分について（図10のSt81）以降の処理を再帰的に繰り返すことを行う。この時、生成中の決定木（99）に選択された特徴量を付け加えることで決定木を成長させる。

【0011】以上の処理における特徴量の選択は、つぎのようにして行う。すなわち、相互情報量計算部96により、与えられたデータSの持つ情報量I(S)と、すべての特徴量A_iについて、その特徴量がどの値である*

$$I(S) = - \sum_{k=1}^n P_k(S) \log_2 P_k(S)$$

$$E(A_i, S) = \sum_{S_j \in A_i \text{ による } S \text{ の分割}} \frac{|S_j|}{|S|} I(S_j)$$

である。

【0013】なお、これについては、「参考文献2」（Quinlan, J. R., "Induction of Decision Trees", Machine Learning Vol. 1, 1986.）に詳しい。この時、もし特徴量中に連続値をとるものがあつたならば（図10のSt84）、連続値特徴量記号化装置95によって、連続値を区間分割することによって記号的特徴量に変換し（図10のSt86）、記号的特徴量と同様に扱う。

【0014】以下では、連続値特徴量記号化装置95において用いられる公知の区間分割アルゴリズムの例について説明する。図11は、連続値特徴量記号化装置95において用いられる公知の区間分割アルゴリズムの例の構成を示したものであり、図12は、連続値の区間を分割する処理（図10のSt86）において用いられる公知の区間分割アルゴリズムの例における処理の流れを示したものである。

【0015】学習データ集合／連続値特徴量入力部71には、決定木学習用データと、区間を分割すべき連続値特徴量とが入力される（図12のSt3.0）。入力された特徴量をAとする。

【0016】次に、学習データ分類部72が、入力された決定木学習用データ79を、その分類クラスによって分類する（図12のSt31）。次に、分類結果序列化装置73が、分類された学習用データ80の各分類結果を、それぞれ特徴量Aの値の平均値により順序付ける（図12のSt32）。序列化された分類済み学習用データ81に対し、境界値決定部74は、隣合った分類結果w₁、w₂に対して、以下の式で決められる境界値Bを計算する。ここで、m_iは分類結果w_iの特徴量Aの平均を、d_iは分類結果w_iの特徴量Aの値の分散を表す。

【0017】

$$B = (m_2 d_1 + m_1 d_2) / (d_1 + d_2)$$

こうして、すべての分類結果に対し境界値を割り当てたなら、再分割必要性判定部75が、再度分割すべき分類

4

*かによってデータを分割した後の各部分の情報量の和E(A_i, S)、さらにその差gain(A_i, S) = I(S) - E(A_i, S)を計算し（St85）、差gain(A_i, S)が最も大きい特徴量を選ぶ（St87）ことによって行う。ここで、分類クラスをC₀, C₁, ..., C_nとし、上記与えられたデータS中でのクラスC_iの出現頻度をP_i(S)としたとき、

【0012】

【数1】

結果があるかどうかを判定する。これは、分類結果w_iに対し、w_i中のデータで、w_iに割り当てられた境界値内に特徴量Aの値が属さないものの割合が、予め決められた割合（cluster rate）より大きいかどうかを判定することにより行う（図12のSt34）。

【0018】その結果、もし大きいなら、例題分割部76が、その分類結果を2つに分割する（図12のSt35）。分割は、例えば、後述するk-means アルゴリズムを、k=2として用いる。

【0019】ここで、一つでも分割された分類結果があつたならば、図12のSt32以降を繰り返す（図12のSt36）。もしなかったならば、区間記号化装置77が、分割された各区間に記号を付与する（図12のSt37）。そして、記号化特徴量出力部78が、特徴量Aを記号化した特徴量を出力する（図12のSt38）。

【0020】なお、これについては、「参考文献1」（荒木大、小島昌一、"決定木学習における数値データの区間分割"、第5回人口知能学会大会論文集、1991.）に詳しい。

【0021】図8は、図7に示す事例データから、上記定められた割合の設定値（prune rate）を70%、与えられたクラスタ誤り率のレート（cluster rate）を40%として上記決定木学習装置を用いて導いた決定木である。

【0022】

【発明が解決しようとする課題】以上述べた従来の技術では、以下のような問題点がある。従来の決定木学習装置においては、事例データ中に分類クラスが与えられることを前提としており、しかもそれは離散的な値をとるものでなければならない。従って、ある特徴量にしたがって分類を行うような決定木を求めたいが、その特徴量が連続値をとる場合には、事前にオペレータ等が区間分割等を行って、離散な分類クラスを定義してやる必要があつた。

【0023】例えば、図3に示すような事例データが与

5

えられ、分類対象の特徴量N0 およびN2 の値に従って分類を行う決定木を求めたいとするならば、たとえば特徴量N0 に対して {620未満、620以上640未満、640以上660未満、660以上680未満、680以上} と云うような区間分割を行い、この区間分割によって得られたそれぞれの区間に、N2 の値であるAの場合とBの場合とで例えば、{C1, C2, C3, C4, C5} と云うような具合に分類クラスの名前を与えた後に、事例データを決定木学習装置に入力する必要がある。図7に示す事例データは、図3に示すものに対してこのような操作を人手によって行い、得られたものを示したものである。

【0024】しかし、人手によるこのような作業は、量が少ない場合は良いが、量が膨大になったり、分類が複雑になると手に負えなくなる。従って、この作業を自動化する必要がある。

【0025】本発明の目的とするところは、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定し、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようにした決定木学習装置を提供することにある。

【0026】

【課題を解決するための手段】上記目的を達成するため、本発明はつぎのように構成する。すなわち、事例の特徴を表す第1の特徴量と、分類用の属性を示す第2の特徴量との少なくとも2種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習生成装置として、与えられた複数の事例データを順次取り込み、その第1の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第1の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成した。

【0027】

【作用】このような構成において、複数の事例データを与えると、決定木生成手段は当該与えられた複数の事例データを順次取り込み、その第1の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第1の特徴値(特徴量)を分類するために予め定めた分類の粒度の値以下となるまで、予め設定されたクラスタ数分に分割して分類クラスとすると共に、各隣接する分

6

類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する。そして、決定木生成終了制御手段は、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了させる。

【0028】本発明は、事例の特徴を表す第1の特徴量と、分類用の属性を示す第2の特徴量との少なくとも2種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則である決定木を、事例データ集合を与えることで生成する決定木学習装置であり、連続値をとる特徴量と属性によってデータの分類を行う場合の決定木導出における各特徴量選択の段階でクラス分け(クラスタリング)を行い、適切な区間をこの特徴量に対して設定する。クラスタリング・アルゴリズムに対するパラメータであるクラスタ数は、ユーザが必要とする分類の精度(分類の粒度)を与えることによって自動的に決定する。決定木導出は、データ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度より小さくなるか、あるいは、その集合の、クラスタリングによって与えられた分類クラスの中でのクラス分け誤り率が設定レート(prune rate)に収まるか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に収まると終了させるようにした。

【0029】このように、連続値を属性に対応して分類する決定木を作成しようとする場合、分類の精度(分類の粒度)と、特徴値と分類すべき属性の値からなる学習用の複数のサンプルデータを与えることによって当該サンプルデータから該サンプルデータの属性別最大差を前記分類の粒度に従った分類クラスにクラスタリングし、その分類クラスの特徴値境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成することから分類クラスを自動的に定義できるので、連続値属性に対する分類木の作成が完全に自動化される。

【0030】従って本発明によれば、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習装置を提供することができる。

【0031】

【実施例】以下、本発明の実施例を図面に基づいて説明する。図1は、本発明の構成を示すものであり、図2は本発明全体の処理の流れを示すものである。図1において、51は決定木学習用データ入力部、52は決定木生

7

成終了判定部、53は決定木生成部、54は特徴量選択部、55は連続特徴量記号化装置、56は相互情報量計算部、57は決定木出力部、58は決定木、59は生成中の決定木、60はクラスタリング部、61はパラメータ入力部である。

【0032】これらのうち、決定木学習用データ入力部51はいくつの特徴量からなる事例のデータを入力するものであり、パラメータ入力部61は分類の対象にする属性名やその分類に必要とされる精度（分類の粒度（prec））、決定木生成終了判定に用いる同一の分類クラスに分類されたデータの分類誤り率を示す分類誤り率設定値（prune rate）、そして、与えられたクラスタ誤り率のレート（clusterrate）などのパラメータ・データを入力するためのものである。

【0033】また、決定木生成終了判定部52は、与えられた訓練例の属性Nに属する枝の取り得る値の最大値と最小値との差（diff）が、分類の粒度（prec）以下であるかどうかを判断し、差（diff）が分類の粒度prec以下であって、すべての再帰的繰返しについて同じことが言えている場合に、全ての再帰の分枝が終了のときは終了と判断し、全ての再帰の分枝が終了していなければその枝は終りと判定してつぎの枝の処理を実行させる指示を行う機能を有する。

【0034】決定木生成部53は、特徴量選択部54によって選択された特徴量がどの値であるかによって、与えられた訓練例を複数の部分に分割し、それぞれの部分について（図2のSt21）以降の処理を再帰的に繰返すことを実施するものであり、連続特徴量記号化装置55は相互情報量計算部56が求める特徴量中に連続値をとるものがあつた時、パラメータ・データである与えられたクラスタ誤り率のレート（cluster rate）を用いてこのレートを満たすことができるならば、前記連続値を区間分割すると云つた処理を行うものである。連続特徴量記号化装置55の構成は図11に示した公知の構成で良く、処理の流れは図12に示した公知のアルゴリズムを利用する。

【0035】また、相互情報量計算部56は、与えられた訓練例集合Sの持つ情報量 $I(S)$ と、全ての特徴量 A_i について、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和 $E(A_i, S)$ 、さらにその差 $gain(A_i, S) = I(S) - E(A_i, S)$ を計算するものであり、特徴量選択部54はこの結果から、上記差 $gain(A_i, S)$ が最も大きい特徴量を選ぶと云つた処理を行うものである。

【0036】決定木出力部57は決定木生成終了判定部52が全ての再帰の分枝終了と判断した際に、決定木生成部53が生成して生成中の決定木データ59として保存した当該決定木データ59を最終的な決定木データ58として出力するものである。クラスタリング部60は訓練例の属性Nの値の集合に対し、「diff/prec」個の

8

クラスタに分割するクラスタリングを行うものである。

【0037】つぎに上記構成の本装置の作用を図2のフローチャートを参照して説明する。いくつかの特徴量からなる事例（以下、訓練例と呼ぶ）が紙、または磁気テープなどに表形式に記録されているとする。例えば図3である。

【0038】図3は、N0、N1、N2と云う3つの特徴量を持った訓練例である。この訓練例のデータは、キーボード入力、ネットワークによるオンライン入力、情報伝達の媒体である磁気テープからの読取り入力等のかたちで決定木学習用データ入力部51より入力される（図2のSt20）。また、パラメータ入力部61からは、分類の対象にする属性名（与えられた事例データの、どの特徴量に基づいて分類を行うかを示すもの）、その分類に必要とされる精度（分類の粒度（prec））、前述の、決定木生成終了判定のために用いる同一の分類クラスに分類されたデータの分類誤り率の割合を示す分類誤り率設定値（prune rate）、連続値特徴量記号化装置55で用いられるクラスタ誤り率のレート（cluster rate）などのパラメータ・データが入力される。

【0039】決定木学習用データ入力部51から訓練例が入力されると（図2のSt20）、決定木生成終了判定部52によって、与えられた訓練例の属性Nの値の最大値と最小値との差（diff）が分類の粒度であるprec以下であるか否かが判断される（図2のSt30）。その結果、差（diff）が分類の粒度prec以下であって、すべての再帰的繰返しについてやはり同じことが言えているならば、全ての再帰の分枝が終了と判断し（図2のSt22）、これによって決定木出力部57が決定木のデータ58を出力し（図2のSt23）、決定木を求める処理を終了する。

【0040】そうでないときはクラスタリング部60でのクラスタリング処理に入る。クラスタリング部60では、訓練例の属性Nの値の集合に対し、「diff/prec」個のクラスタに分割すると云つたクラスタリングの処理を行う（図2のSt29）。これには、例えば公知のk-means アルゴリズムを用いる。

【0041】すなわち、クラスタリング部60はクラスタリングの処理をつぎのようにして行う。今、各クラスタを分類のためのクラス（以下分類クラス）と考える。この時、各クラスタの境界値は、そのクラスタの最小値と値の小さい側に隣合ったクラスタの最大値との中間値、およびそのクラスタの最大値と値の大きい側に隣合ったクラスタの最小値との中間値とする。最小（最大）のクラスタの小さい（大きい）側の境界値は設定されないものとする。例えば、

{667}, {674, 681}, {688}

と云うクラスタが出来た時、{667}のクラスタは「無限小から“670.5”以下」と云う名前の分類クラスと考え、{674, 681}のクラスタは「“67

9

0.5”より大きく、“684.5”以下」と云う名前の分類クラスと考えるのである。

【0042】クラスタリング部60によって分類クラスが定義された後は、前述した公知の決定木学習装置と同様に、特徴量を選択することによって決定木を生成する。すなわち、これはつぎのようにして行う。

【0043】クラスタリング部60による分類クラスの定義が終了すると、決定木生成終了判定部52での処理に入り、まず、決定木生成終了判定部52において、分類クラスに分類された事例データの分類誤り率が、同一の分類クラスに分類されたデータの分類誤り率設定値 (prune rate) 以上を占めるかどうか判断される (図2のSt21)。その結果、もし分類誤り率設定値 (prune rate) 以上を占めていて、すべての再帰的繰り返しについてやはり同じことが言えているならば (図2のSt22)、決定木出力部57が決定木 (58) を出力し (St23)、決定木を求める処理を終了する。

【0044】図2のステップSt21の判定の結果、もし分類誤り率設定値 (prune rate) 以上を占めていないならば、決定木生成部53での処理に移る。そして、決定木生成部53では、特徴量選択部54によって選択された特徴量がどの値であるかによって、与えられた訓練例を分割してそれぞれの部分について (図2のSt21) 以降の処理を再帰的に繰り返すことを実施し、これによって、当該訓練例の取り得る枝を複数に分岐し、それぞれの部分についての特徴量を求める。

【0045】この時、生成中の決定木59の枝に対して、上述の選択された特徴量を付け加えることで決定木生成部53は決定木を成長させる。以上の処理における特徴量の選択は、相互情報量計算部56の計算結果をもとに特徴量選択部54により行われる。すなわち、相互情報量計算部56では与えられた訓練例集合Sの持つ情報量I(S)と、全ての特徴量Aiについて、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和E(Ai, S)、さらにその差gain(Ai, S) = I(S) - E(Ai, S)を計算するので (図2のSt25)、特徴量選択部54はこの相互情報量計算部56が求めたもののうち、上記差gain(Ai, S)が最も大きい特徴量を選ぶ (図2のSt27)。これによって特徴量の選択が成される。

【0046】この時、もし特徴量中に連続値をとるものがあつたならば (図2のSt24)、連続値特徴量記号化装置55によって、連続値を区間に分割することによって記号的特徴量と同様に扱う。この分割処理は図12のフローチャートに従う。

【0047】以下では、クラスタリング部60において用いられる公知のクラスタリングアルゴリズムの例について説明する。図4は、クラスタリング部60において用いられる公知のクラスタリング・アルゴリズム例の構成を示したものであり、図5は、分類対象特徴量のクラ

10

スタリング (図2のSt29) において用いられる公知のクラスタリング・アルゴリズム例における処理の流れを示したものである。

【0048】クラスタリング部60によるクラスタリング処理は図4に示すように、クラスタリングされるべきデータと、パラメータ(k)とを入力するためにあるデータ/k入力部11から、まず、クラスタリングされるべきデータと、いくつかのクラスタリングに分けるかと云うパラメータ(k)が入力されることにより開始される (図5のSt41)。データ/k入力部11から入力されたこれらのデータは、データ記憶部12に保持される。

【0049】このデータ記憶部12に保持されたデータは、クラスタリング実行部17と平均値初期化部13に与えられる。すると、平均値初期化部13はクラスタの平均を、データ記憶部12に保存されたデータの先頭から、相異なるk個を取ってその平均値を求め、各クラスタの平均値保持部15に更新記憶させることによって各クラスタの平均値の初期化をする (図5のSt42)。

【0050】一方、クラスタリング実行部17は、データ記憶部12に記憶されている各データを、それと最も近い平均値を持つクラスタへと分類する (図5のSt43 (クラスタリング処理))。

【0051】そして、その結果はクラスタリング結果としてクラスタリング結果保存部18保存する。ついで各クラスタの平均値保持部15に保持されている平均値DAV-NEWを前回の平均値DAV-OLDとして前回の各クラスタ平均値保持部14へとコピーした後、クラスタリング結果保存部18に対して、平均値計算部20が各クラスタの平均値を計算し (図5のSt44)、この計算により得られた平均値を新しい平均値DAV-NEWとして各クラスタの平均値保持部15に保存する。

【0052】ここで、クラスタ平均比較部16が、各クラスタの平均値保持部15に保存された新しい平均値DAV-NEWと、前回の各クラスタ平均値保持部14に保持されている前回の平均値DAV-OLDとを比較して両者が等しいか否かを判断する (図5のSt45)。

【0053】クラスタ平均比較部16による当該判断の結果、両者が等しければ、出力部19はクラスタリング結果保存部18からクラスタリング結果を出力して (図5のSt46) 終了する。

【0054】ステップSt46の比較の結果、両者が等しくなければステップSt43以下の処理を繰り返す。以上がクラスタリング部60におけるクラスタリング処理操作の詳細である。

【0055】図6は、図3の学習用データに対して、分類対象特徴量をN0、分類の粒度 (prec) を“10”、与えられたクラスタ誤り率のレート (cluster rate) を“40%”、分類誤り率設定値 (prune rate) を“99%”として本発明を用いて生成した決定木である。

11

【0056】図6の決定木は、「特徴量N1が“621.5”以下であり、特徴量N2がAであるなら、特徴量N0の値は“670.5”から“684.5”の間である。そうでなく(N2がBで)、N1が以下なら、N0の値は“656.5”から“670.5”の間である。そうでなく…」と云う知識を表す。

【0057】従来の方法においては、決定木の“葉”に現れる値(の範囲)は、学習システムの外側でユーザが与える必要があったが、本発明によれば、決定木の“葉”に現れる値は、与えられた精度に対して適切な値(適切な値の範囲)となるように自動的に決定できる。

【0058】要するに、本発明は、数値または記号で表現された幾つかの特徴量と、それに対してオペレータ等が事前にデータを吟味して与える分類結果を事例とし、多数の事例を分類装置の例題として与えることによって、特徴量と分類結果の間の一般法則を見付け出し、新たな事例が与えられた時に、その事例の持つ特徴量を評価して分類結果を出力すると云った処理をする学習装置において、事例の特徴を表す第1の特徴量と、分類用の属性を示す第2の特徴量との少なくとも2種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習装置として、与えられた複数の事例データを順次取り込み、その第1の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第1の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成したものである。

【0059】そして、このような構成において、複数の事例データを与えると、決定木生成手段は当該与えられた複数の事例データを順次取り込み、その第1の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第1の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成し、そして、決定木生成終了制御手段は、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、

12

定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了させるものである。

【0060】本発明は、事例の特徴を表す第1の特徴量と、分類用の属性を示す第2の特徴量との少なくとも2種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則である決定木を、事例データ集合を与えることで生成する決定木学習装置であり、連続値をとる特徴量と属性によってデータの分類を行う場合の決定木導出における各特徴量選択の段階でクラス分け(クラスタリング)を行い、適切な区間をこの特徴量に対して設定する。そして、クラスタリングに必要なパラメータであるクラスタ数は、ユーザが必要とする分類の精度(分類の粒度)を与えることによって自動的に決定するようにしており、また、決定木導出は、データ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度より小さくなるか、あるいは、その集合の、クラスタリングによって与えられた分類クラスの中でのクラス分け誤り率が設定レート(prune rate)に収まれば終了させるようにした。

【0061】そしてこのように、連続値を属性に対応して分類する決定木を作成しようとする場合、分類の精度(分類の粒度)と、特徴値と分類すべき属性の値からなる学習用の複数のサンプルデータを与えることによって当該サンプルデータから該サンプルデータの属性別最大差を前記分類の粒度に従った分類クラスにクラスタリングし、その分類クラスの特徴値境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する方式を採用したことにより、分類クラスを自動的に定義できるので、連続値属性に対する分類木の作成を完全に自動化することができるようになる。

【0062】従って本発明によれば、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習生成装置が得られる。なお、本発明は上述した実施例に限定するものではなく、その要旨を変更しない範囲内で適宜変形して実施し得るものである。

【0063】

【発明の効果】以上詳述したように本発明によれば、連続値属性に対する分類木を作ろうとする場合、クラスタリングによって分類クラスが自動的に定義されるので、分類木の作成が完全に自動化されるなど、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習生成装

置を提供できる。

【図面の簡単な説明】

【図 1】本発明の実施例を説明するための図であって、本発明の一実施例の構成図。

【図 2】本発明の実施例を説明するための図であって、本発明の一実施例における装置の動きを示す流れ図。

【図 3】本発明の実施例を説明するための図であって、本発明の実施例で入力として用いる訓練例のデータを示す図。

【図 4】本発明の実施例を説明するための図であって、本発明の実施例におけるクラスタリング部 60 の詳しい構成図。

【図 5】本発明の実施例を説明するための図であって、本発明の一実施例におけるクラスタリング部 60 の装置の詳しい動きを示す流れ図。

【図 6】本発明の実施例を説明するための図であって、図 3 に示すデータと、所定の学習用パラメータに対し、本発明の一実施例である装置が生成した決定木を示す図。

【図 7】従来技術を説明するための図であって、公知の決定木学習装置が扱える事例データの例を示す図。

【図 8】従来技術を説明するための図であって、図 7 の事例データに対して公知の決定木学習装置が導いた決定*

*木を示す図。

【図 9】従来技術を説明するための図であって、公知の決定木学習装置の構成を示すブロック。

【図 10】従来技術を説明するための図であって、公知の決定木学習装置の処理の動きを表す流れ図。

【図 11】従来および本発明の実施例において使用される連続値特徴量記号化装置 95、55 の詳しい構成図。

【図 12】従来および本発明の実施例において使用される連続値特徴量記号化装置 95、55 の装置の詳しい動きを示す流れ図。

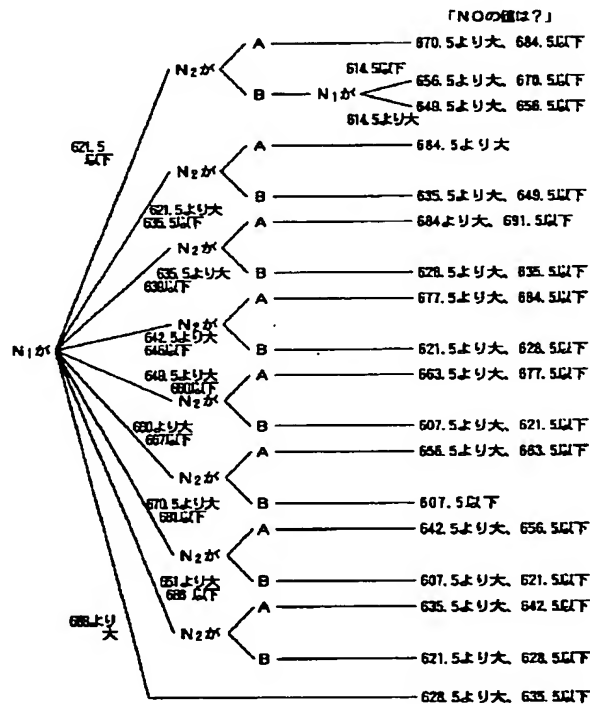
【符号の説明】

- 51…決定木学習用データ入力部
- 52…決定木生成終了判定部
- 53…決定木生成部
- 54…特徴量選択部
- 55…連続特徴量記号化装置
- 56…相互情報量計算部
- 57…決定木出力部
- 58…決定木
- 59…生成中の決定木
- 60…クラスタリング部
- 61…パラメータ入力部

【図 3】

事例番号	N ₀	N ₁	N ₂
1	688	639	A
2	681	646	A
3	674	653	A
4	667	660	A
5	660	667	A
6	653	674	A
7	646	681	A
8	639	688	A
9	632	695	A
10	625	688	B
11	618	681	B
12	611	674	B
13	604	667	B
14	611	660	B
15	618	653	B
16	625	646	B
17	632	639	B
18	639	632	B
19	646	625	B
20	653	618	B
21	660	611	B
22	667	604	B
23	674	611	A
24	681	618	A
25	688	625	A
26	695	632	A
27	688	639	A
28	681	646	A

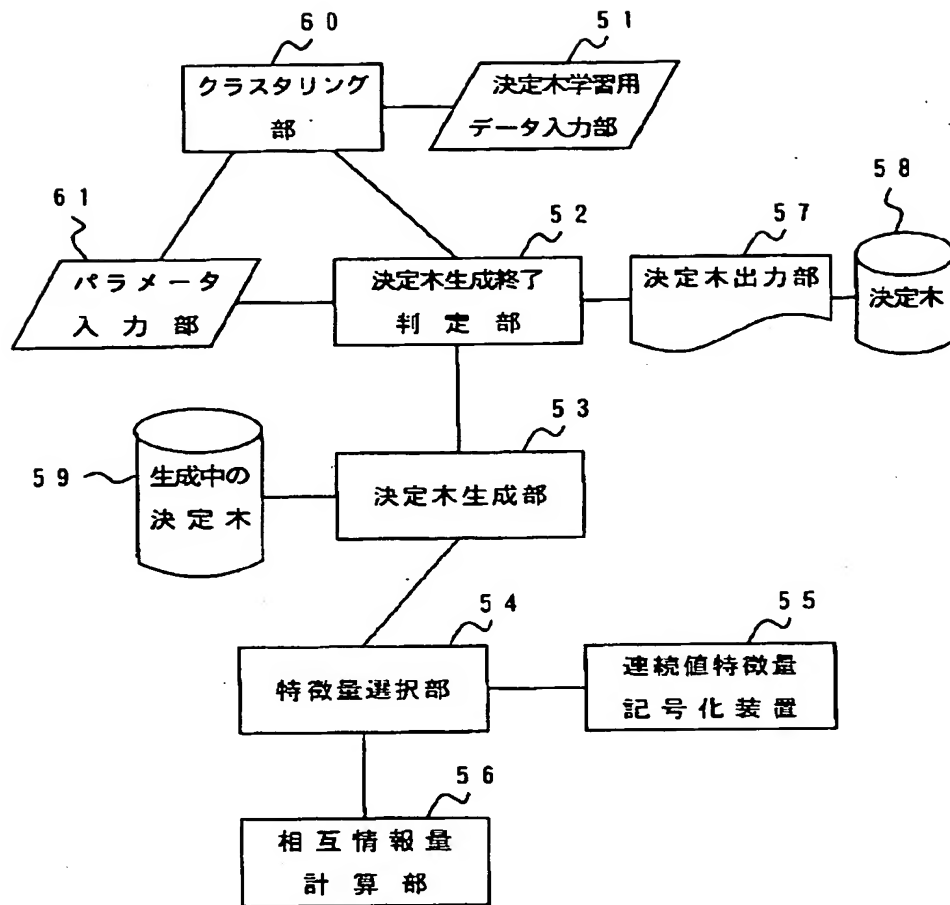
【図 6】



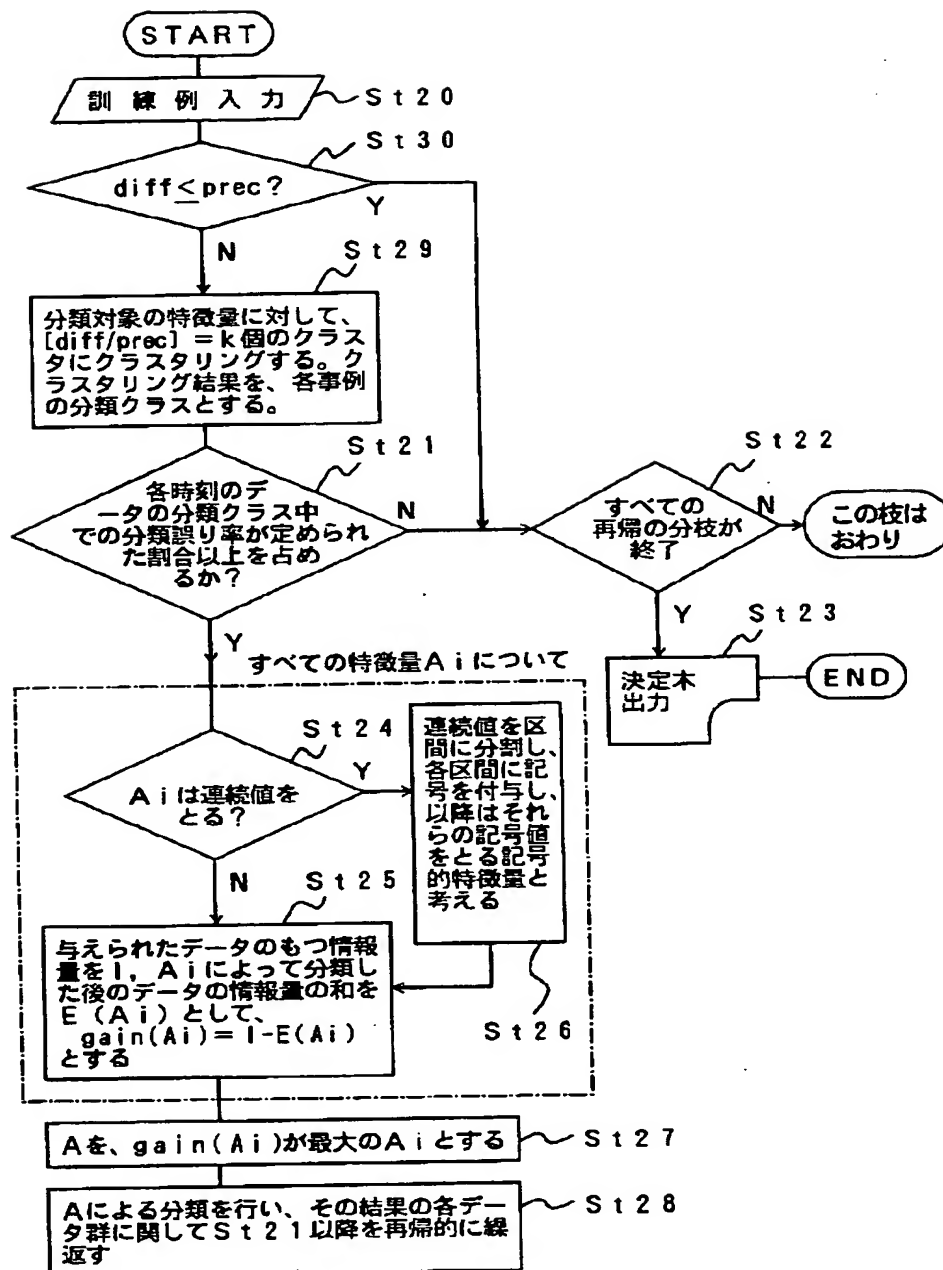
【図 7】

事例番号	分類クラス	N ₁	N ₂
1	C ₁	639	A
2	C ₂	646	A
3	C ₃	653	A
4	C ₄	660	A
5	C ₅	667	A
6	C ₆	674	A
7	C ₇	681	A
8	C ₈	688	A
9	C ₉	695	A
10	C ₁₀	688	B
11	C ₁₁	681	B
12	C ₁₂	674	B
13	C ₁₃	667	B
14	C ₁₄	660	B
15	C ₁₅	653	B
16	C ₁₆	646	B
17	C ₁₇	639	B
18	C ₁₈	632	B
19	C ₁₉	625	B
20	C ₂₀	618	B
21	C ₂₁	611	B
22	C ₂₂	604	B
23	C ₂₃	611	A
24	C ₂₄	618	A
25	C ₂₅	625	A
26	C ₂₆	632	A
27	C ₂₇	639	A
28	C ₂₈	646	A

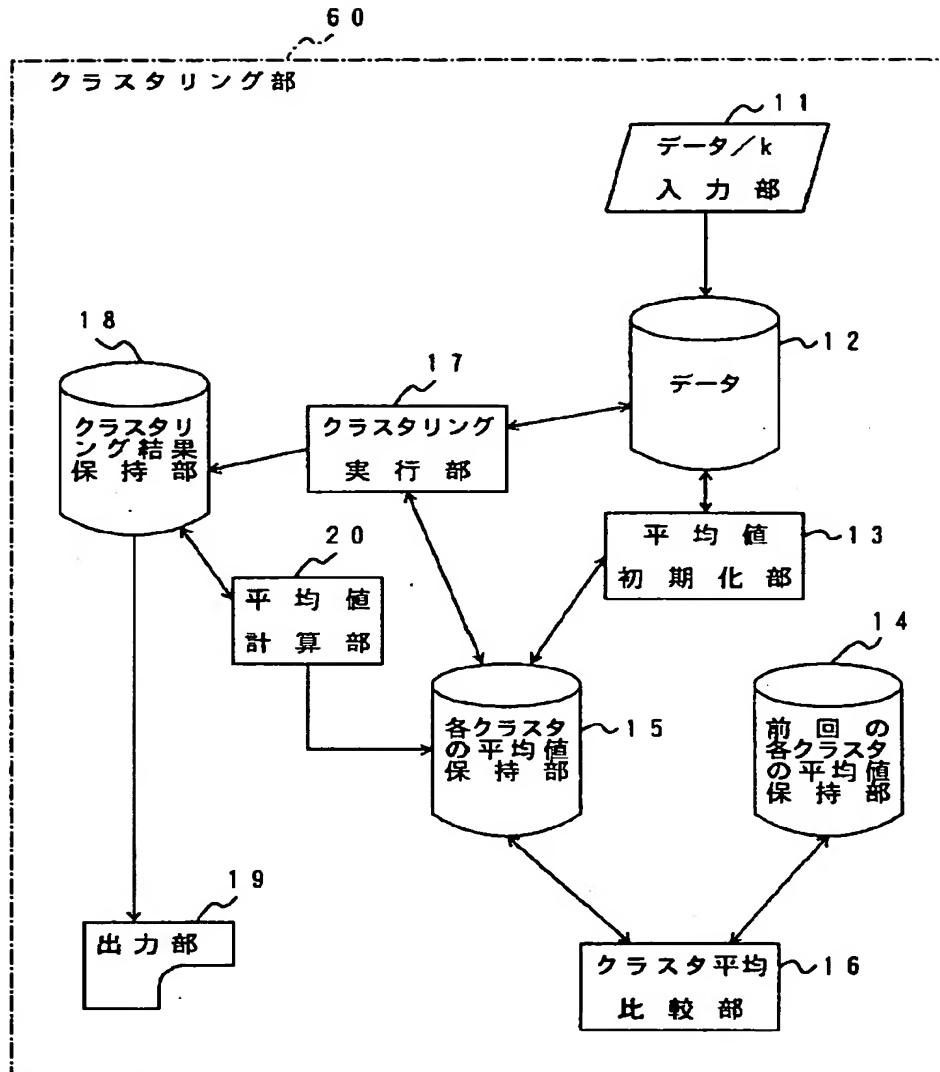
【図 1】



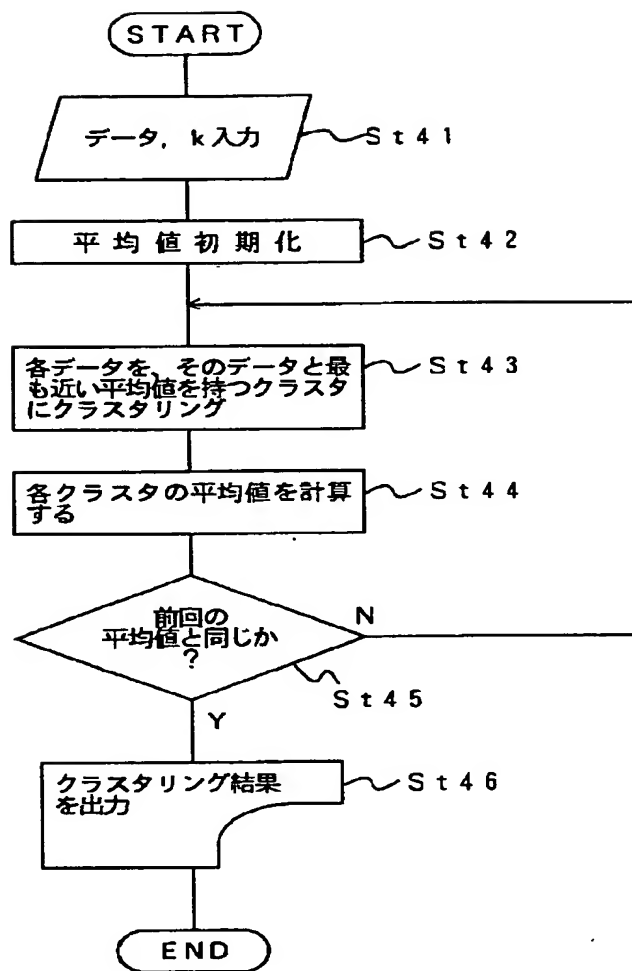
【圖 2】



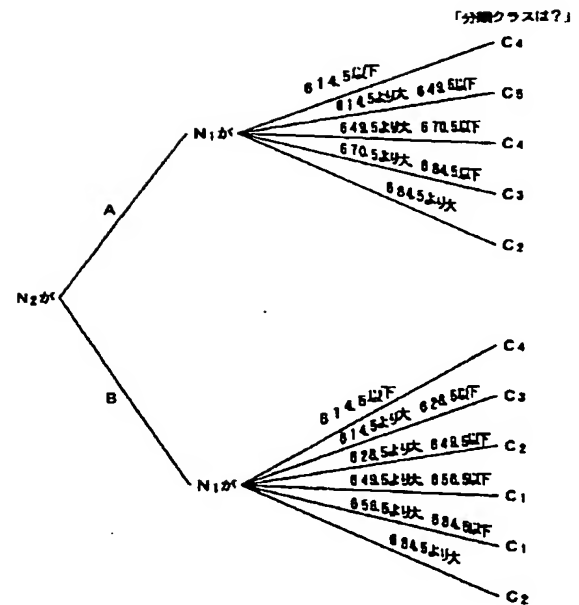
【図4】



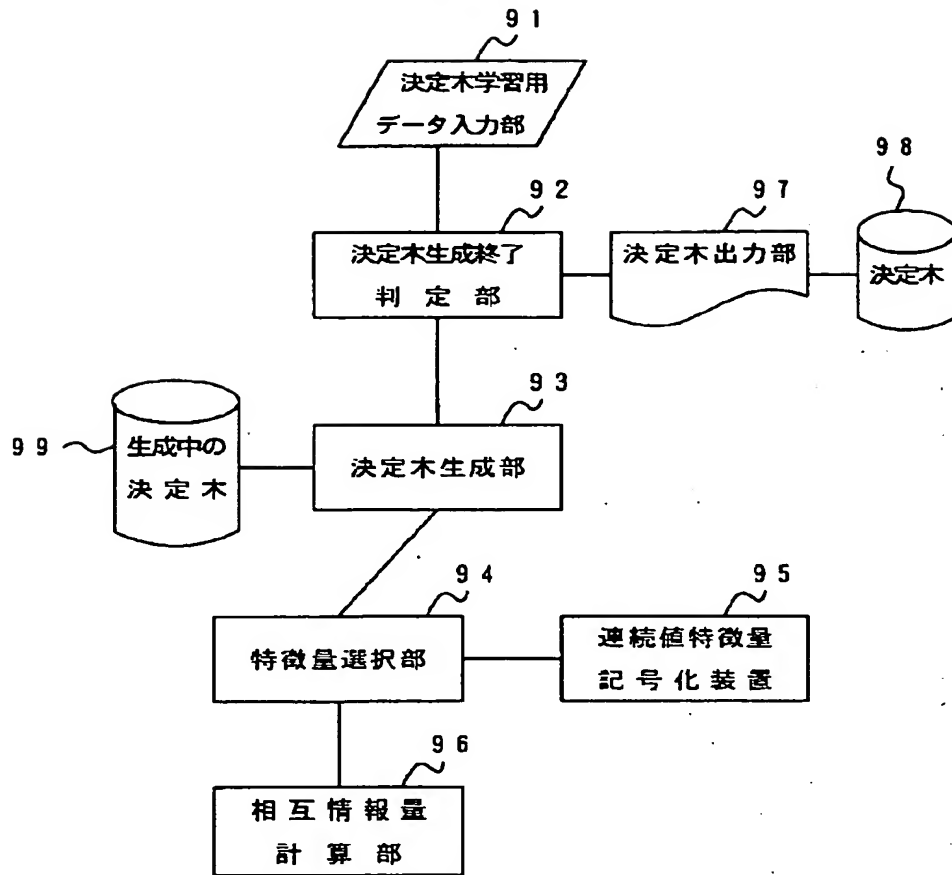
【図 5】



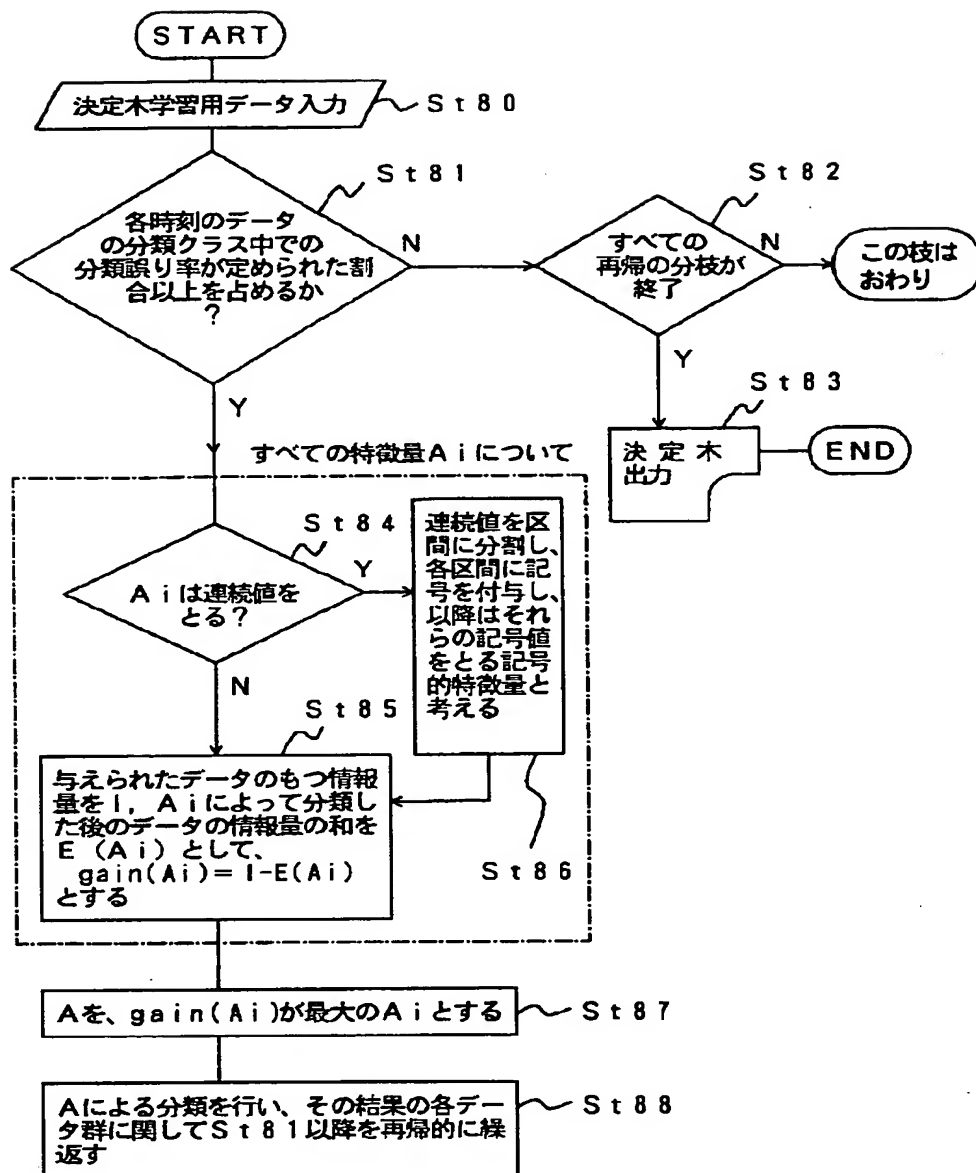
【図 8】



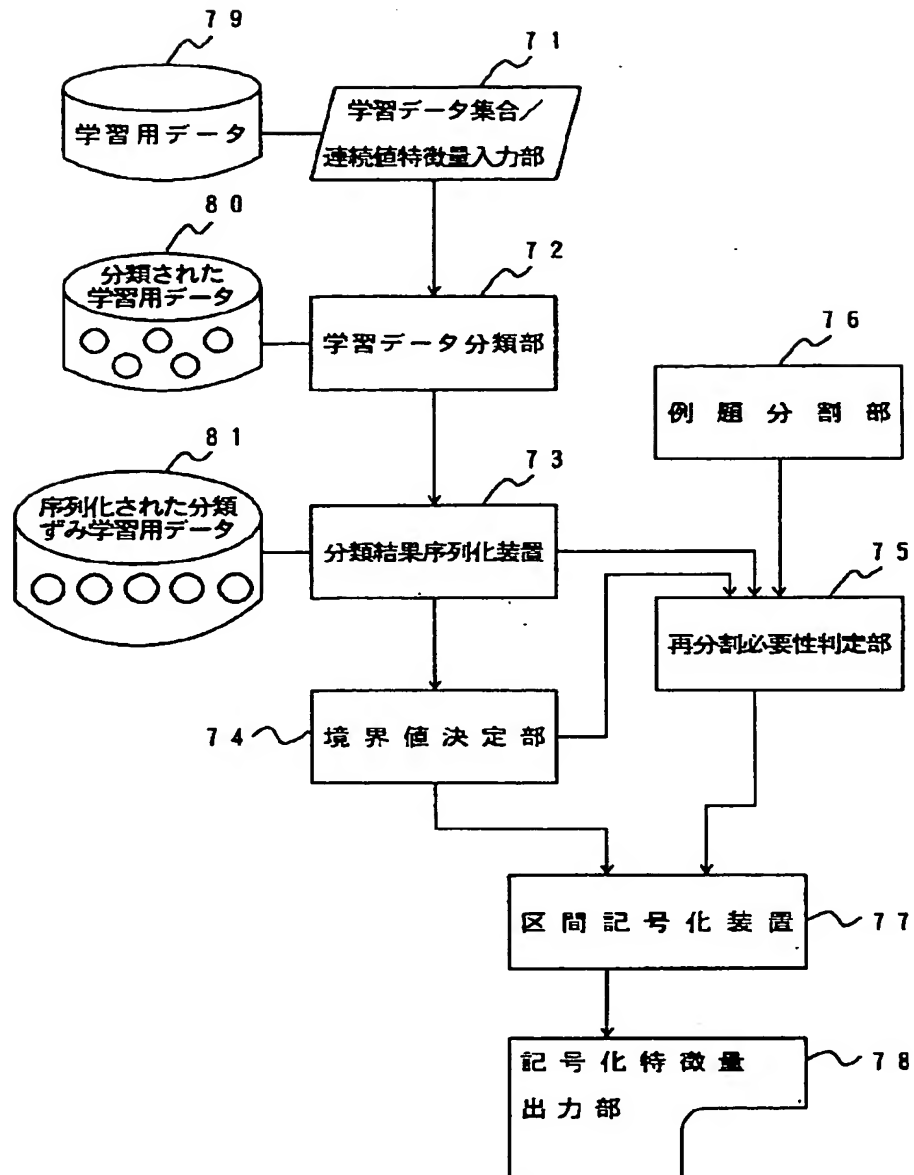
【図 9】



【図 10】



【図 11】



【図12】

